Galia Z. Amram
Tracy Racicot Hucke
Asst. Federal Public Defenders
214 W. Lincolnway Ste. 31A
Cheyenne WY 82001
307-772-2781
California State Bar #250551
Wyoming State Bar 7-4880
Galia_Amram@fd.org
Tracy_Hucke@fd.org

UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF WYOMING

UNITED STATES OF AMERICA,	)	
	)	
Plaintiff,	)	
	)	
v.	)	No. 18-CR-020-SWS
	)	
ARAPAHO JAMES OLDMAN,	)	
	)	
Defendant.	)	

MOTION TO EXCLUDE EXPERT TESTIMONY ON STRMIX-GENERATED DNA INCLUSION OR MATCH STATISTIC UNDER *DAUBERT* AND FEDERAL RULES OF EVIDENCE 702

Table of Contents

I. INTRODUCTION

On April 21, 1992, a 22-year-old woman was raped behind a vacant building in Muncie, Indiana. Based on the description she gave at the scene, police canvassed the area and picked up 35-year-old William Barnhouse. The police took Barnhouse to the scene and stood him next to three police cars. As officers shined flashlights in his face, the victim identified Barnhouse as her attacker.

Barnhouse went to trial in Delaware County Circuit Court in December 1992. A crime lab blood analyst from the Indiana State Police said he was able to "match" genetic markers in the biological evidence found on the woman's jeans and in the rape kit, and that he could not eliminate Barnhouse as the source of the evidence. A hair analyst from the Indiana State Police said that a hair found on the woman's body was a "match" for Barnhouse. In closing argument, the prosecution told the jury that the hair was a "silent witness" against Barnhouse. On December 15, 1992, the jury found Barnhouse guilty but mentally ill of rape and criminal deviant conduct. He was sentenced to 80 years.

Over twenty years later, in 2013, the FBI reported that testimony asserting that microscopic hair comparison could produce a "match" between two hairs was scientifically invalid. A subsequent review of FBI analysts' testimony and reports determined that analysts had provided erroneous testimony or reports in more than 90 percent of the cases reviewed by 2017. In 2016, the Innocence Project and the Wrongful Conviction Clinic at Indiana University sought DNA testing of the sperm in the vaginal swabs in the rape kit and on the sperm found on the victim's jeans. The testing identified the same male DNA profile in the rape kit and on the jeans,

and excluded Barnhouse as the source of the biological evidence.[1]

Mr. Barnnouse's story is not unusual. Misapplication of forensic science is the second most common contributing factor to wrongful convictions, found in nearly half (45%) of DNA exoneration cases.[2] In response to the spate of exonerations of innocent defendants wrongfully convicted based in part on forensic-science evidence, the government undertook studies to determine the validity and reliability of forms of forensic evidence and testimony based upon them.[3] The most recent study occurred in 2015 when President Obama asked the President's Council of Advisors on Science and Technology (PCAST) to consider whether there were steps that could be taken on the scientific side to strengthen the forensic-science disciplines and ensure the validity of forensic evidence used in the Nation's legal system. Exh. A (PCAST Report), at 14. The report issued by the PCAST Commission reviewed a number of types of forensic evidence including, as is relevant to this case, DNA analysis. The PCAST Commission found DNA analysis of single-source and simple-mixture samples to be reliable. As to complex-mixture samples, which are described below, the PCAST commission found:

> at present, studies have established the foundational validity of some objective methods under limited circumstances (specifically, a three-person mixture in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture) but that substantially more evidence is needed to establish foundational validity across broader settings.

Exh. A at 21.

As the Court is aware, Arapaho Oldman is charged in the above-captioned case

[1] https://www.innocenceproject.org/cases/william-barnhouse/ (last visited Oct. 29, 2018).

[2] DNA Exonerations in the U.S., INNOCENCE PROJECT, http://www.innocenceproject.org/dnaexonerations-in-the-united-states/ (last visited Oct. 29, 2018).

[3] A Congressionally-mandated study released in 2009 by the National Research Council, Strengthening Forensic Science in the United States: A Path Forward, was particularly critical of weaknesses in the scientific underpinnings of a number of the forensic disciplines routinely used in the criminal justice system. Exh. A (PCAST Report), at 14.

with First Degree Murder. In Oldman's case, just as in Barnhouse's case, law enforcement personnel scoured the crime scene for items to send to the crime lab. Seventy items were sent to the lab for assorted forms of forensic testing, including DNA analysis on 23 items. Exh. B (FBI's DNA Report). For 21 of those items, the analysis met the requirements of the PCAST commission and Arapaho Oldman was excluded as a contributor. For two of those items, items 12 and 53, the FBI violated the PCAST commission's strictures and Oldman was potentially inculpated. Item 12, a swab from the basement stairs where the body of the victim was found, was a mixture of four or more people where the minor contributor, purportedly Oldman, was less than 20% of the mixture. Item 53, a stain on Oldman's shirt, was also a mixture of four or more people and the analysis showed that Oldman's co-defendant, Whiteplume was inconclusive as a possible contributor. Despite the PCAST commission's determination that DNA analysis of complex mixtures of four or more people is not foundationally valid, the FBI did the testing anyway and the result is the only forensic evidence tying Oldman to the crime. All the other forensic testing – including all the other foundationally valid DNA testing – excluded Arapaho Oldman.

And so this Court is faced with a decision: should it allow risky, untested forensic evidence in at trial or, now that we know the long, sordid history that bad forensic science has played in convictions of innocent people, should the Court use its gatekeeping power under *Daubert* and Fed.R.Evid. 702 and 703 to exclude it.

II. BACKGROUND: THE CRIME AND THE INVESTIGATION

On December 11, 2017, the government filed a criminal complaint charging Oldman with one count of First Degree Murder for the killing of Charles Dodge III sometime between

November 22, 2017 and November 30, 2017. (Docket No. 1). The Complaint was based on facts outlined in a statement by Federal Bureau of Investigation Special Agent Christine Coble. Coble stated that on November 30, 2017, Wind River Police Department found the remains of Dodge in a crawlspace in the basement at 331 Great Plains Road in Arapahoe, WY. One of the owners of the home said he had been told that an unknown, intoxicated woman told him that Oldman had killed a man and buried the body in the crawlspace.

Coble entered the basement and saw blood spatter, transfers, swipes and droplets in "numerous locations." There were shoe impressions on the stairs in blood, as well as pooling of blood, and items soaked in blood, near the entry to the crawlspace. The body was found covered in the crawlspace and due to lividity, Coble believed, and the coroner later confirmed, that Dodge was not killed in the position in which he was found, but rather moved there after death. Dodge suffered blunt force trauma to the head, sharp force trauma to the neck, as well as mutilation of the face. Coble began investigating who was responsible for Dodge's death.

Early on in the investigation, a key witness identified Oldman as the prime suspect. Witness #1 said that s/he was drinking in the basement of 331 Great Plains before Thanksgiving with Dodge, Oldman, Whiteplume and Witness #2. S/he said that when the group ran out of vodka, Oldman demanded some from Dodge. When Dodge refused, Oldman became enraged and began to beat Dodge with closed fists. Witness #1 claims s/he threw herself on top of Dodge to protect him but Oldman punched him/her twice in the back and then pulled him/her off of Dodge. Oldman then began beating Dodge with a metal wrench, hitting him repeatedly in the head, torso and limbs. Witness #1 said Whiteplume also hit Dodge, kicking him twice in the head without shoes on. Witness #1 further claimed Oldman beat Dodge until Dodge was unresponsive and then dragged Dodge into the crawlspace. Witness #1 stated that Oldman

placed Dodge in the crawlspace by himself and no one helped him. Witness #1 then explained that s/he checked on Dodge twice throughout the night and he was still alive but unresponsive. At daybreak, s/he left the residence and hitchhiked to Riverton. (Docket No. 1.) Based largely on the statement of Witness #1, Oldman was charged with First Degree Murder.[4]

The government continued to investigate while the criminal case was pending. As part of the investigation, Coble sent 70 items from the crime scene to the FBI crime lab in Quantico, Virginia. (Docket No. 62). This included items Coble believed could be a murder weapon including a dented, bloody soup can, a metal bar and a rock,[5] as well as clothing from Dodge, Oldman, Whiteplume, and Witness #1, and swabs from throughout the crime scene. These items were tested for both DNA and fingerprints. Pictures of shoeprint impressions were also taken and sent to the FBI crime lab. The government sought and obtained two continuances in order to obtain the results from the crime lab. (Docket No. 62 and 96).

The crime lab results were not consistent with either Witness #1 or Whiteplume's statements. Exh. B (DNA Report). Oldman was excluded as a contributor to a bloody sweatshirt and jacket found outside the crawlspace, a bloody shoe behind the couch, and a swab of the collar of Dodge's sweatshirt and jacket (though Whiteplume was included as a contributor to the collar of Dodge's sweatshirt and inconclusive as to the jacket). *Id.* Oldman was also excluded as a contributor to swabs from the cooler, wall, the step next to furnace, and multiple steps at and around the crawlspace entry. *Id.* Oldman was further excluded as a contributor to the DNA

[4] There were other witness statements including from co-defendant Whiteplume and from people who claim to have been told what happened by Oldman or Whiteplume. In addition, one witness stated s/he saw blood on Oldman's shorts and another stated s/he saw blood on Oldman's hands. However, the only witnesses who say they were present and witnessed the beating are Witness #1 and Whiteplume. Whiteplume claimed Oldman beat Dodge to death and he (Whiteplume) only hit Dodge in the chest twice and helped put the body in the crawlspace. Amram Decl., at ¶ 3. Based on these statements, Whiteplume was initially charged with accessory after the fact. (Docket No. 23).

[5] Statements made by Coble to the crime lab personnel, provided in discovery, state that she believed these items could be the murder weapon. Amram Decl., at ¶ 4.

found on the bloody, dented can and the bloody rock. No DNA found was on the metal bar. *Id.* at 6. There was a possible bloodstain found on Oldman's bandana but Dodge, Twiss and Whiteplume were all excluded as contributors to that.[6] Exh. B at 8. No usable fingerprints were found at the scene and so there is no fingerprint evidence tying Oldman to the murder either. Amram Decl., at ¶ 4.

The only forensic evidence possibly tying Oldman to the crime or its known participants was two items. The first, Item #12, is a swab from the steps of the basement at 331 Great Plains. Exh. B at 5. The swab was found to contain a mixture of four people (though, as discussed below, it very well may be more) and, as will be explained in detail below, a software program called STRMix was used to analyze the probability that a person could be included as a contributor to the mixture. *Id. See also,* Exh. E (lab documents re Item 12). Oldman, whose DNA was deemed to constitute 15% of the mixture, Exh. E at 7, was included as a potential contributor with a likelihood ratio of 1.0 million (meaning that the DNA results are one million times more likely if he is a contributor than if he is not). Exh. B at 5. For item 53, Oldman's shirt, a possible bloodstain was also found to contain a mixture of four people (though, as discussed below, it very well may be more) and STRMix was again used to analyze the probability that a person could be included as a contributor to the mixture. Dodge was excluded as a contributor but Whiteplume was inconclusive. Exh. B at 8; Exh. F (lab documents re Item 53).

Additional information uncovered during the investigation casts doubt on the credibility of Witness #1 account as Oldman as the killer. Oldman passed a polygraph. Amram Decl. at ¶ 5. Numerous other potential witnesses, including Whiteplume, failed a polygraph. Amram

[6] A man named Lonestar Addison, who is not charged in the case, was included as a contributor. Exh. B at 8.

Decl., at ¶ 6. After initial interviews denying culpability and a failed polygraph by Lamebull, Jori Lamebull and Monty Tabaho both confessed to helping Whiteplume hide the body in the crawlspace and Tabaho confessed to attempting to cut off Dodge's head on Whiteplume's orders. Amram Decl., at 6. Both Lamebull and Tabaho said Oldman was not present during this time. *Id.* In addition, Whiteplume's DNA was included as a possible contributor to a mixture found on Dodge's throat. Oldman was excluded. Exh. B. Based on Lamebull and Tabaho's confessions, they are now charged as accessories after the fact and Whiteplume is now charged with First Degree murder. (Docket No. 108 [superseding indictment]).

III. HOW DNA WORKS

Deoxyribonucleic acid, or DNA, is a double-stranded molecule that coils to form the characteristic double helix, and is found in all cells possessing a nucleus.[7] John Butler, *Fundamentals of Forensic DNA Typing* ("*Fundamentals*"), 19 (2010). Forensic DNA typing examines certain locations, or loci, on the DNA strand. The DNA typing technique at issue in this case is short tandem repeat (STR) testing. STR typing measures how many times a short piece of DNA repeats at each of the tested loci; the number of repeats is known as an allele. *Id* at 148. An individual's genetic type, or profile, is the compilation of his or her alleles at each locus tested. At each locus, an individual possesses two alleles: one allele inherited from each biological parent. *Id* at 25. Thus, an individual's DNA profile is simply a list of two numbers per locus examined. An individual can inherit the same allele—*i.e.* same number of repeats—at a locus from his or her biological parents (*e.g.* 12, 12). This means the individual is a homozygote at that locus. Alternatively, an individual can inherit two different alleles—two

[7] Most cells, with the exception of red blood cells, possess nuclei. When it is in the nucleus, DNA is tightly packaged into two sets of 23 chromosomes; one set of 23 chromosomes is inherited from each parent. Sperm and egg cells possess only one set of 23 chromosomes each; when they unite, the resulting embryo possesses the full set of 46 chromosomes. *Fundamentals* at 23.

different numbers—at a locus from his or her parents (*e.g.* 12, 16). This means they are a heterozygote at that location.

The DNA testing process proceeds via a series of steps: extraction, quantitation, amplification, analysis of genetic data, evaluation, comparison and interpretation. The first step in generating a DNA profile from a sample is extraction, where the analyst attempts to isolate the DNA and separate it from all other cellular material and debris. *Id.* at 99. After extraction of the DNA, the sample is quantitated, *i.e.*, the total amount of DNA present in the sample is estimated. *Id*. at 114. Based on the estimated amount of DNA present, some portion of the extracted DNA is then amplified. Amplification is a process by which DNA is copied at targeted locations (i.e. loci) many times over, generating on the order of a billion copies.[8] *Id.* at 125-26. During the amplification process, the targeted DNA may not amplify if there is an insufficient amount of DNA to start with[9], or if it is degraded (i.e. broken into pieces due to environmental exposure or other stressors), or if there are inhibitors (such as some fabric dyes or excess salts) present in the sample. *Id*. at 68. When targeted DNA does not amplify, that genetic information is lost in downstream steps; this loss of genetic information is known as allelic dropout, a concept discussed further below. *Id*. at 222.

The post-amplification sample consists of large numbers of only the copied alleles, which can then be separated on an instrument called a genetic analyzer so that each allele can be distinguished and then recorded. *Id.* at 175. The result of this process is a series of peaks on a graph, called an electropherogram. *Id*. at 194. The analyst interprets the electropherogram,

[8] Amplification is conducted via a technique called polymerase chain reaction, commonly notated as PCR.
[9] Quantitation gives a preliminary estimate of whether the amount of DNA in the extract falls into this low level range. However, a seemingly sufficient total amount of DNA may be comprised of low levels of DNA from multiple contributors; this is not something that can be discerned from the quantitation step, which does not distinguish between contributors but rather reports the total amount of DNA present.

generating a genetic profile for the evidence sample. Part of the interpretation process involves determining whether the peaks present represent "real" DNA or artifacts of the testing process. Each "real" DNA peak corresponds to an allele present in the sample and the height of each peak roughly corresponds to how much of that allele is present (*i.e.* a taller peak indicates more of a particular allele present). When testing a single source evidence sample (i.e. a sample originating from one individual), two peaks of roughly equivalent height should be observed at each locus where the contributing individual is a heterozygote (i.e. possesses 2 different alleles). At loci where the contributor is a homozygote (i.e. possesses two of the same allele), one, relatively high peak should be observed, because the two alleles "stack" on top of one another.

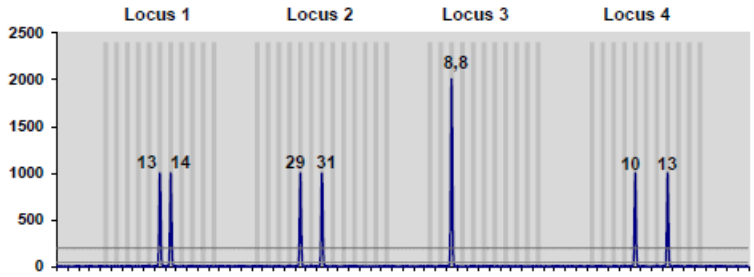

Figure 1. Electropherogram showing ideal, single-source DNA data at four hypothetical loci. Note that at Locus 3, where the DNA contributor is a homozygote (possesses two "8" alleles), his two alleles "stack" on top of one another, resulting in a single peak on the electropherogram. At each of the other three loci, where the contributor is a heterozygote (i.e. possesses two different alleles), two peaks are observed. Figure from Butler, *Advanced Topics in Forensic DNA Typing: Interpretation*, 11, Fig. 1.5 (2014).

After the evidence sample is interpreted, the analyst then compares the resulting profile to the profile that the analyst developed from the reference sample(s).

If the analyst determines that one of the reference profiles "match" or "cannot be excluded from" the evidence profile, the analyst calculates a rarity statistic to contextualize the significance of the match or inclusion. Statistical calculations are the second step in the interpretation process, giving the trier of fact a means of assessing the possibility that the

inclusion is coincidental. "The statistical calculation step is the pivotal element of DNA analysis, for the evidence means nothing without a determination of the statistical significance of a match of DNA patterns." *People v. Barney* (1992) 8 Cal.App.4th 798, 817. This is because "it would not be scientifically justifiable to speak of a match as proof of identity in the absence of underlying data that permit some reasonable estimate of how rare the matching characteristics actually are." National Research Council, *The Evaluation of Forensic DNA Evidence, Committee on DNA Forensic Science: An Update* (1996) [*NRC II*][10]; *see also* National Research Council, *DNA Technology in Forensic Science* (1992) [*NRC I*][11] at 74 ("To say that two patterns match, without providing any scientifically valid estimate (or at least, an upper bound) of the frequency with which such matches might occur by chance, is meaningless"). The Scientific Working Group for DNA Analysis Methods, or SWGDAM,[12] created to provide discipline-wide guidelines, similarly admonishes that analysts "must perform statistical analysis in support of any inclusion that is determined to be relevant in the context of a case, irrespective of the number of alleles detected and the quantitative value of the statistical analysis." SWGDAM Interpretation Guideline 4.1 (2010).[13]

IV. INTERPRETATION ISSUES WITH COMPLEX DNA MIXTURES

Forensic DNA samples from crime scenes often contain DNA from more than one individual. Such a DNA profile representing two or more contributors is termed a DNA mixture. *Fundamentals*, *supra*, at 320. An analyst knows that they are dealing with a DNA mixture,

[10] *Available at* https://www.nap.edu/catalog/5141/the-evaluation-of-forensic-dna-evidence (accessed November 28, 2018).
[11] *Available at* https://www.nap.edu/read/1866/chapter/1 (accessed November 28, 2018).
[12] SWGDAM is an advisory group convened by the Federal Bureau of Investigation. https://www.swgdam.org/ (accessed November 28, 2018).
[13] *Available at* http://www.forensicdna.com/assets/swgdam_2010.pdf (accessed November 28, 2018).

versus a single source sample, if they observe more than two alleles at two or more loci, or if loci with only two alleles display significant peak height imbalance.[14] John Butler, *Advanced Topics in Forensic DNA Typing: Interpretation* ("*Interpretation*"), 129 (2014). Unlike the straightforward analysis involved in interpreting a high-quality single source DNA profile, mixtures are often ambiguous, and the process of interpreting them can be highly subjective. In particular, mixtures which cannot be resolved into single source components ("indistinguishable" mixtures)[15] involve a great deal of subjective decision making. Studies have shown that subjective interpretation of indistinguishable DNA mixtures can lead to widely divergent results from one analyst to the next, even analysts in the same laboratory applying the same set of protocols. *See* Dror and Hampikian, Subjectivity and bias in forensic DNA mixture interpretation, *Sci. & Justice*, 51(4), 204–208 (2011)[16]; NIST Interlaboratory Mixture Interpretation Study 2013 ("MIX13")(discussed *infra*) and forerunner NIST mixture studies (e.g. MIX05). Two complicating factors in mixture interpretation are: (1) "the potential for allele stacking", and (2) "potential alleles in the stutter position." *Interpretation* at 153.

[14] Two alleles from the same contributor should be roughly the same height, within a degree of tolerance (called a "peak height ratio" (PHR). If the height of two allelic peaks observed at a given locus are not within this predetermined tolerance—i.e. they are "imbalanced"—this is a sign that they actually originate from two people rather than one.

[15] Mixtures may sometimes be resolved or 'deduced' into individual sources based on the relative amounts of DNA contributed by each source; the source contributing more DNA is the 'major contributor', and the source(s) contributing less DNA is the 'minor contributor.' Analysts use the height of allelic peaks on the electropherogram as a proxy for how much DNA is originating from each contributor. Laboratories have specific criteria regarding how much difference they have to observe between peak heights to pull out a major profile; it would be inappropriate to 'eyeball' a mixture to determine whether it impressionistically appears that there is 'enough' of a difference between contributors to deduce a major profile. Some mixtures encountered in casework do not meet these criteria and therefore the mixture must be treated in its totality rather than as individual single source profiles.

[16] In this study, 17 examiners from one government laboratory were provided a mixed DNA profile from a sex assault case and asked to interpret the profile and compare it to a suspect's reference profile. The original case work analyst had determined that the suspect could not be excluded as a contributor to the mixture. The 17 examiners came to a variety of conclusions: 1 concluded "cannot exclude"; 12 "excluded" and 4 deemed the results "inconclusive". Among other things, these results underscore the subjectivity of complex mixture interpretation. *Available at* http://www.scienceandjusticejournal.com/article/S1355-0306(11)00096-7/pdf (accessed November 28, 2018).

a. **The Problem Of Allele Stacking**

As described *supra*, when an individual has two of the same alleles at a locus (*i.e.* is a homozygote), that person's alleles "stack" on top of one another and present as a single peak on the electropherogram. Similarly, when multiple contributors to a DNA mixture possess the same allele at a locus, those alleles also "stack" and present as a single peak. *See, e.g.,* Figure 2, below. This is known as allele stacking or allele sharing. There are two important consequences of allele stacking. One consequence is that "allele sharing makes accurately deducing the number of contributors to a mixture challenging – and the challenge only grows with each additional contributor to a DNA mixture." *Interpretation* at 169. If an analyst cannot accurately determine how many contributors there may be in a mixture, the analyst cannot reliably interpret the mixture. Studies have shown that, because of allelic stacking, more than 75% of known four-person mixtures would be misclassified as two- or three- person mixtures based on the maximum number of alleles detected at any given locus. Paoletti et al., Empirical Analysis of the STR Profiles Resulting from Conceptual Mixtures, J Forensic Sci, 1361-66 (2005)[17]. Inaccurate interpretation of the mixture impacts whether an individual is included or excluded as a potential contributor to the mixture, as well as the associated statistical analysis. *Interpretation.* at 335 ("some of these genotype combinations may not fit a reasonable interpretation of the data" depending on the actual number of contributors present).

[17] A copy of which is available upon request.

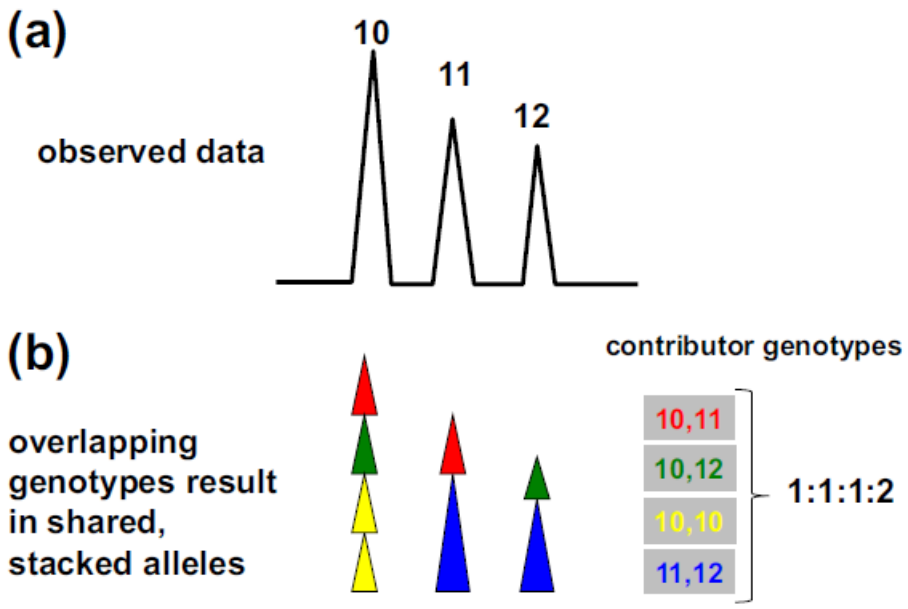

Figure 2. Hypothetical mixture which (a) exhibits only three alleles at a locus (and is thus suggestive of a two person mixture), (b) is actually comprised of three low level contributors plus a single higher level contributor, whose alleles stack on top of one another. Figure from *Interpretation*, at 160, Fig. 7.1.

A second consequence of allele stacking is that it diminishes the utility of the stochastic threshold as a means of determining whether allelic dropout has occurred. *Interpretation* at 163 ("the potential of allelic stacking, especially with more than two contributors, can limit the usefulness of a stochastic threshold."). When the stochastic threshold cannot be effectively utilized, interpretation and statistical analysis are not reliable. Reliance on a stochastic threshold without considering the possibility of allelic dropout may result in a false inclusion or exclusion.

A stochastic threshold is a Y-axis value on the electropherogram (measured in relative fluorescence units, or RFUs). The value is established by the laboratory's internal validation studies.[18] Data below the stochastic threshold is in the "potential 'danger zone' of unreliable

[18] The stochastic threshold value(s) should be a part of the lab's written protocols as determined by the lab's validation.

results." John M. Butler and Carolyn R. Hill, *Scientific Issues with Analysis of Low Amounts of DNA* (2010).[19] When peaks from the evidence sample fall below the stochastic threshold at a locus, there is a risk that allelic dropout—or loss of genetic data—is occurring at that locus. Specifically, allelic dropout occurs when only one of a DNA contributor's two alleles at a given locus is detected by the DNA typing process.[20] This is a common problem associated with low template DNA analysis. As described *supra,* allelic dropout can happen when an individual's DNA is present in low levels, is degraded, or is inhibited. Stochastic thresholds are important because seeing an allele below the stochastic threshold at a given locus is a "warning indicator" that the partner allele (*i.e.* the second allele of the pair) may have dropped out. *Id.* at 163-64. Dropout of a partner allele could lead an analyst to detect a "false homozygote." *Id.* For example, if the true contributor of an evidentiary DNA sample possesses an 8 and a 12 allele at a given locus, but due to allelic dropout only the 8 allele is detected, an individual who is homozygous for the 8 allele (*i.e.* possesses two 8 alleles) could be falsely implicated, while the true contributor could be falsely excluded. *See, e.g., infra* at 18-21 (discussion of Case 5 in MIX13 study). While the stochastic threshold serves some purpose, in that DNA data that is unambiguously in the stochastic range (i.e. below the stochastic threshold) is clearly at risk of being incomplete, DNA data that rises above the stochastic threshold is not necessarily safe. This is especially true with complex DNA mixtures, due to the potential for contributors to share alleles (i.e. allele stacking).

Allele stacking makes over-reliance on using the stochastic threshold to determine whether or not drop-out has occurred particularly dangerous with mixtures. "Just because allelic

[19] https://www.promega.com/resources/profiles-in-dna/2010/scientific-issues-with-analysis-of-low-amounts-of-dna/ (accessed November 28, 2018).
[20] There can also be loss of both alleles at a given location, which is called locus dropout.

peaks at a locus are above an established stochastic threshold does not mean that no allele drop-out has occurred in a complex mixture." *Interpretation* at 163-64. Allele stacking can falsely elevate a peak at a particular locus above the stochastic threshold. When two or more sub-threshold alleles stack on top of each other, they may present as a peak that surpasses the stochastic threshold, which in turn may give the false impression that the DNA at that locus is free from the risk of allelic dropout and can be confidently interpreted. In reality, however, each contributor to the falsely elevated peak is in the stochastic (dropout) zone. *See, e.g.,* Figure 2 (showing peaks from multiple low-level contributors stacking upon one another and presenting as three relatively tall peaks). "The concept of a stochastic threshold can become meaningless in complex mixtures due to the potential for allele stacking." *Interpretation* at 94-95; *id.* at 177 ("stochastic thresholds often lose their value and meaning when allele sharing is possible with three or more contributors to a DNA mixture").

While some level of allelic stacking will occur in any DNA mixture, there is no objective way to determine whether allelic stacking is occurring at any given locus in an indistinguishable DNA mixture profile because there is no way to tell whether an observed peak comes from one contributor, or actually is the combined low level (*i.e.* sub-stochastic) contributions of two or more individuals. *See, e.g., supra,* Figure 2 (the 10 allele demonstrates how an above-threshold peak can originate from a combination of two or more individuals whose individual contributions are below the stochastic threshold). Analysts typically make educated guesses based on other information present in the DNA profile because there are no objective guidelines or protocols to guide the analyst's determination that allele stacking is or is not occurring. These

educated guesses may or may not lead to accurate conclusions. The more complex the mixture the more difficult it is to make a reliable educated guess.[21] Ultimately, and unavoidably, "allele drop-out and potential allele sharing from multiple contributors lead to greater uncertainty in the specific genotype combinations that can be reliably assumed." *Interpretation* at 177. And, as Dr. Butler has unambiguously warned, "[w]hen there is a high degree of interpretation uncertainty from an evidentiary sample, it makes little sense to try and draw conclusions . . . and expect those conclusions to be reliable." *Id.*[22]

b. The Problem Of Stutter

Another significant source of uncertainty in mixture interpretation is distinguishing real alleles from artifacts, particularly in an extremely common artifact known as "stutter." Stutter is a by-product of the amplification (*i.e.* copying) step in the DNA testing process, and typically results in a small peak one repeat less than its parent "true allelic" peak (*e.g.* the process would produce a smaller "stutter" peak in the 7 allele position when there is a true 8 allele at that locus). "Because stutter products are the same length as actual allele PCR products, it can be challenging to determine whether a small peak is a real allele from a minor contributor[23] of the original sample or a stutter product of an adjacent allele created during the PCR amplification process."

---

[21] This is why most labs have a "complexity rule," that precludes interpretation of samples that are too complex either because of the number of contributors or the potential for allelic drop-out.

[22] Allelic dropout is not simply a theoretical possibility. It is "ever-present" and a "real issue faced with complex mixtures," because "[s]ensitive DNA detection technology has the potential to outpace reliable interpretation." *Interpretation* at 174, 177. "If a laboratory desires to develop appropriate protocols that will enable reliable interpretation of DNA from low-level DNA or mixtures involving three or more contributors, then validation studies need to be performed with known samples that mimic the amounts of DNA and complexity of profiles where stochastic effects and allele dropout are expected." *Id.* at 164. The 2017 SWGDAM Interpretation Guidelines require internal validation studies to establish the stochastic threshold, while acknowledging that reliance on the stochastic threshold may not be appropriate in mixture samples where allele sharing is possible. SWGDAM Interpretation Guidelines (2017) 1.7, 1.7.1, and 1.7.1.3. The FBI's validation studies applicable to Items #12 and #53 are discussed in Exhibit H.

[23] A minor contributor is simply an individual contributing a smaller amount of DNA (which will appear on the electropherogram as smaller peaks) relative to other contributors to a DNA mixture.

Id. at 76. When there are one or more minor contributors present whose alleles are similar in height to the stutter peaks, this task is not just "challenging," it is impossible. *Id*. at 58-59. When there is an optimal amount of DNA present, stutter peaks tend not to exceed a certain height relative to the associated parent "true allelic" peak. However, the fact that a low-level peak is adjacent to a larger peak does not necessarily mean that it is stutter. *Id.* at 142 ("It is not always possible to exclude stutter since they are allelic products and differ from their associated allele by a single repeat unit").

For complex mixtures, stutter is even more problematic. Not only does it become impossible to distinguish real DNA from stutter, but stutter peaks can stack in exactly the same way real allelic peaks do. *Id.* at 71. Thus, stutter can stack on a sub-threshold allelic peak and present as a peak that artificially surpasses the stochastic threshold. Moreover, with a mixture containing one or more low level contributors, "higher levels of stochastic variation can lead to more variability in peak height ratios of heterozygotes and more significant stutter products." *Id.* at 160. In other words, when there are low-level DNA contributors present in a mixture, stutter peak heights can exceed expected values (i.e. the values set by validation studies) and be confused with real allelic peaks. In fact, in low template DNA samples stutter peaks may often exceed their "parent" peak making distinguishing a true peak from an artifact impossible. Therefore, with complex mixtures, "This variation leads to a lower confidence in appropriately allocating allele pairs into individual contributor genotypes with complex mixtures". *Id.*

c. The Problem Of Accurately Determining The Number Of Contributors To Complex Mixtures

"The number of contributors **always matters during the interpretation** of mixture evidence." *Interpretation* at 335 (emphasis in original). For example, if a mixture profile

contains no locus with more than four alleles, and the analyst interprets this as a two-person mixture, he may believe that there is no evidence of allelic dropout, and all loci are safe to use in calculating the probability statistic. However, if the same mixture is assumed to have three or more contributors, dropout is likely (and more likely with more contributors). *See, e.g.*, Discussion of NIST inter-laboratory study, Case 5, i*nfra*. The more loci excluded from the probability calculation, the higher the chance an innocent person could erroneously be included as a potential contributor to the limited data that is left. As Dr. Butler points out, "there is a reduced ability to exclude innocent people when loci are dropped out from consideration in the evidence-to-known comparison due to the possibility of allele drop-out." *Interpretation* at 335. There are at least two studies that have shown the difficulty of correctly determining the number of contributors to a complex mixture: the MIX13 study and the Coble/Bright study.

i. The MIX13 Study

In 2013, the National Institute of Standards and Technology (NIST) conducted an inter-laboratory study specifically designed to measure consistency in DNA mixture interpretation across the U.S.[24] NIST Interlaboratory Mixture Interpretation Study 2013 ("MIX13").[25] In particular, NIST was interested in seeing if the 2010 SWGDAM guidelines' recommendation that all laboratories implement a stochastic threshold resolved the wide variation in mixture interpretation practices within and between laboratories that had been observed in earlier NIST mixture studies (e.g. MIX05).

[24] "Exploratory interlaboratory tests are one way the forensic community uses to demonstrate that the methods used in one's own laboratory are reproducible in another laboratory and comparable results are generated by these laboratories. These results are essential to demonstrate consistency in results from multiple laboratories . . . ." *Fundamentals* at 303.

[25] The NIST site containing the details of study design and electronic data that was interpreted by participants is available at http://www.cstl.nist.gov/strbase/interlab/MIX13.htm (accessed November 30, 2018).

The NIST MIX13 study was the largest study of its kind, broadly assessing the accuracy, reproducibility, and repeatability of mixture interpretations among and across laboratories. Analysts from one hundred and eight laboratories took part, and forty-six states had at least one laboratory participate; the participants were from a mix of federal, state, and local labs. As one of the study's leading authors has noted, "[d]ue to the number of laboratories responding and the federal, state, and local coverage obtained, this MIX13 interlaboratory study can be assumed to provide a reasonable representation of current U.S. forensic DNA lab procedures across the community." Dr. Michael Coble, Interpretation Errors Detected in a NIST Interlaboratory Study on DNA Mixture Interpretation in the U.S. (July 22, 2015)("MIX13 Interpretation Errors").[26]

The results of the MIX13 study exposed a disturbing number of errors and showed that, following issuance of the 2010 SWGDAM guidelines, "mixture interpretation is still all over the place." *Id.* at 37. All participants were provided with the same five mock case scenarios and the same set of five evidentiary DNA profiles to interpret, one for each case. Ground truth was known by the study's authors for each case used in the study. As a result, the study authors were able to assess whether false exclusions or false inclusions were made. Two of the five cases (Case 1 and Case 4) involved two-person mixtures, and participants were provided with reference samples for a victim and a suspect who were true contributors to the mixture; for these cases, it was not possible to make a false positive error. However, the study showed that even for two-person mixtures, the calculation of statistics varied widely, with some laboratories improperly using loci with alleles below the stochastic threshold. *Id.* at 26. For each of the three

[26] Dr. Coble's powerpoint discussing the results of the MIX13 study is available online through NIST at https://www.nist.gov/sites/default/files/documents/2016/11/22/interpretation_errors_detected_in_a_nist_interlab_study_on_dna_mixture_interpretation_in_the_us_mix13.coble_.crim1_.pdf (accessed November 30, 2018).

cases where false positives were possible (Cases 2, 3, and 5), because non-contributors were provided among the reference samples, both false inclusions (implicating an innocent person) and false exclusions (excluding the true contributor) were made. *Id.* at 16, 22, 23.

Case 5 involved a four-person mixture which, because of significant allele stacking, could be erroneously interpreted as a two-person mixture. *Id.* at 29, 30; Figure 3, *infra*. ***Sixty-nine percent of participants falsely included an innocent individual in this mixture***. *Id.* at 34. If inconclusive opinions are removed from the total, ***92% of participants making a conclusive determination made a false positive error, implicating an innocent individual***.[27] Notably, all of the peaks in this mixture profile were well above stochastic threshold; unless the participants considered the possibility of allelic stacking, it would not be apparent that allelic dropout might [be occurring] have occurred.

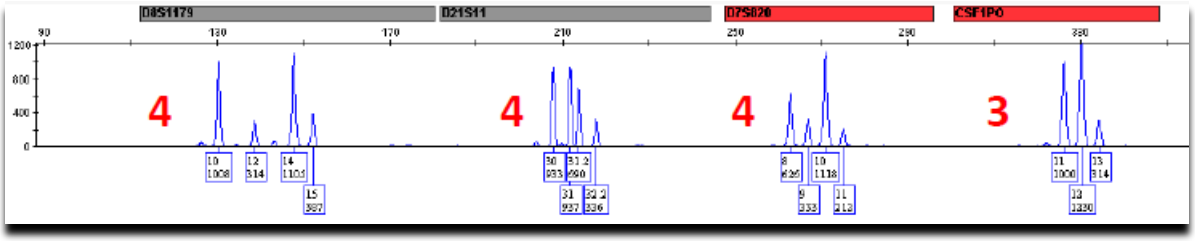

Figure 3. Portion of the mixture electropherogram from Case 5 in the MIX13 study. Note that while this mixture is actually composed of four contributors, the fact that no more than 4 alleles are detected at any locus could give the erroneous impression that the mixture is comprised of two contributors (2 alleles per contributor). The assumption of two contributors (or even three) would cause the analyst to discount the possibility that allelic stacking is bringing peaks above the stochastic threshold, and the related possibility of allelic dropout.

[27] "When reporting a false positive rate to a jury, it is scientifically important to calculate the rate based on the proportion of *conclusive* examinations, rather than just the proportion of all examinations. . . . [C]onsider an extreme case in which a method had been tested 1000 times and found to yield 990 inconclusive results, 10 false positives, and no correct results. It would be misleading to report that the false positive rate was 1 percent (10/10,000 examinations). Rather, one should report that 100 percent of the conclusive results were false positives (10/10 examinations)." PCAST report at 51-52.

ii. The Coble/Bright Study

This year a study was released in which multiple laboratories, including the FBI lab, participated. Exh. H (Declaration of Dan Krane at ¶ 13-14. ["Krane Decl."])[28] Overall, the study showed that labs have become very good at interpreting three-person mixtures and remain very bad at interpreting four or more person mixtures – in large part because labs misidentify five person mixtures as four persons mixtures 58% of the time, and misinterpreted six person mixtures 100% of the time.[29] The study of samples of known origin (lab-made) demonstrated that no lab underestimated the number of contributors to known three-person mixtures. However, approximately 2/3 of evaluated known five-person mixtures and all evaluated known six-person mixtures resulted in underestimates of the true number of contributors. Notably, the majority of known 4-, 5-, and 6-person mixtures were all estimated to contain DNA from four contributors. Exh. H at 14.

[28] It is defense counsel's understanding that one of the authors, T. Moretti, works for the FBI.

[29] J. Bright, R. Richards, M. Kruijver, H. Kelly, C. Mcgovern, A. Magee, A. Mcwhorter, A. Ciecko, B. Peck, C. Baumgartner, C. Buettner, S. Mcwilliams, C. Mckenna, C. Gallacher, B. Mallinder, D. Wright, D. Johnson, D. Catella, E. Lien, C. O'Connor, G. Duncan, J. Bundy, J. Echard, J. Lowe, J. Stewart, K. Corrado, S. Gentile, M. Kaplan, M. Hassler, N. Mcdonald, P. Hulme, R. H. Oefelein, S. Montpetit, M. Strong, S. Noël, S. Malson, S. Myers, S. Welti, T. Moretti, T. McMahon, T. Grill, T. Kalafut, M. Greer-Ritzheimer, V. Beamer, D. A. Taylor, and J. S. Buckleton, *Internal validation of STRmix TM – A multi laboratory response to PCAST*, Forensic Sci. Int. Genet., vol. 34, no. January, pp. 11–24, 2018.

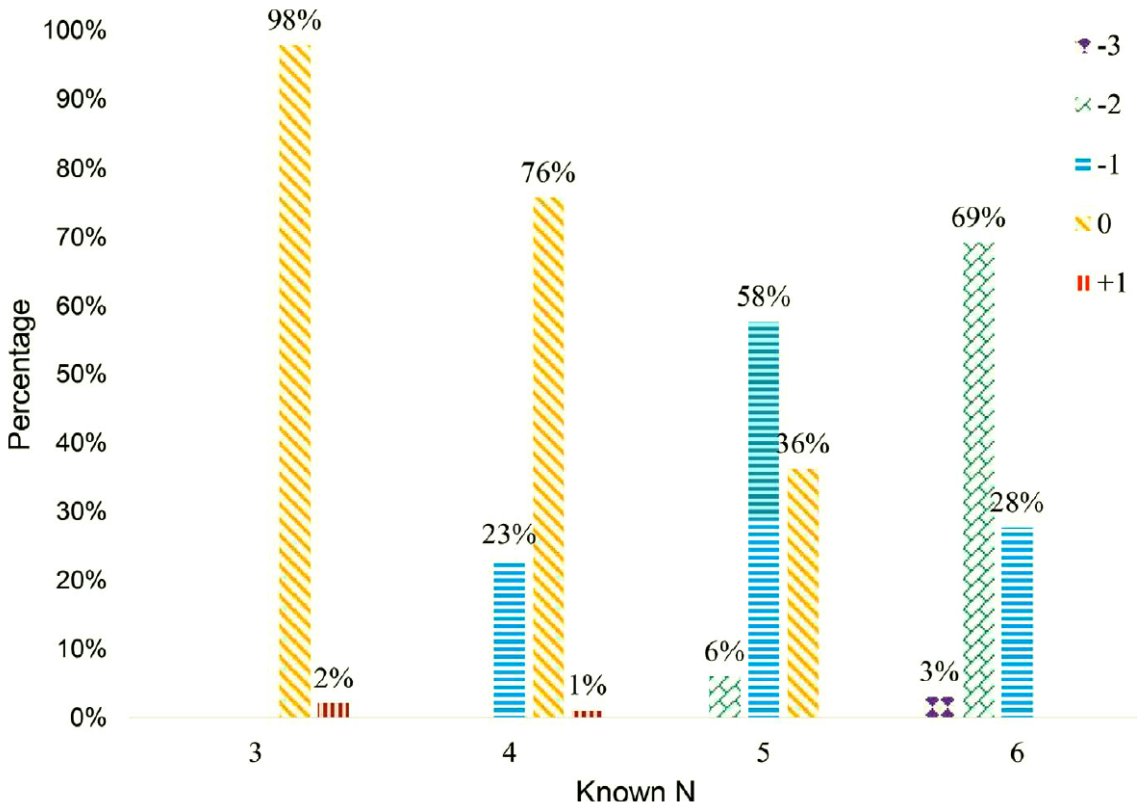

Figure 4 from Bright, et al. – "Plot of percentage of mixtures showing various differences between apparent N and known N against known N. As an example, −1 indicates apparent N was one fewer than known N."

d. Calculating Probabilities For DNA Analysis

As explained above, statistical calculations are the second step in the interpretation process, giving the trier of fact a means of assessing the possibility that a person is included or excluded as a contributor to a DNA sample. As the studies cited above showed, low level DNA samples and DNA mixtures of two or more contributors pose a problem to DNA forensic analysts. In the past analysts dealt with this challenge by calculating statistics concerning the probability of inclusion. But these statistics were general in nature and continue to be the subject of much controversy. *See*, William C. Thompson, Laurence D. Mueller, and Dan E. Krane, *Forensic DNA Statistics: Still*

Controversial in Some Cases, THE CHAMPION, December 2012, 12-23. Recently, labs have begun using software programs to analyze complex DNA mixtures. Probabilistic genotyping software programs are designed to calculate a statistic to contributors of such mixtures when one could not be determined in the past. These programs use biological modeling, statistical theory, computer algorithms, and probability distributions to calculate likelihood ratios (LRs). LRs are the statistic calculated by these probabilistic programs, which reflects the relative probability of a particular finding under alternative theories about its origin. *Id*. In forensic DNA analysis, that LR can be stated as the profile is X amount of times more likely if the defendant and a certain number of other unknown, unrelated contributors contributed to the mixture.

i. What Is STRmix?

STRmix, the probabilistic genotyping software at issue in this case, was developed by the Institute for Environmental Science and Research (ESR) in New Zealand. It uses computer science algorithms to perform "complex" mathematical and statistical calculations. *See* Jo-Anne Bright, Duncan Taylor, et al. "Developmental validation of STRmix, expert software for the interpretation of forensic DNA profiles, Forensic Science Int'l: Genetics"(accepted manuscript), 2016, p. 227 (Hereinafter "Developmental validation of STRmix"). Indeed, STRmix is a software program—it is not used for any of the other steps of DNA analysis described above. *See generally, id.* It is only after a DNA analyst in a laboratory perform the regular steps developing a DNA profile from a sample, that STRmix adds an additional step performed not in a lab by a trained scientist, but instead, by a person sitting at a computer screen, running a complex computer software program. This program is designed to answer the classic question in forensic DNA interpretation: what are the profiles of the contributors to this mixture?

The program relies on analysts to collect the data by reviewing the electopherograms (epg) developed in a case and discarding the peaks below the lab's analytic threshold. *See* Duncan Taylor, Jo-Anne Bright, and John Buckleton, *The Interpretation of Single Source and Mixed DNA Profiles*, Forensic Sci Intl: Genetics 7 519 (2013). Artifacts like pull-up and forward stutter are also removed manually. *Id*. Analysts must also input their determination of the mnumber of contributors to the mixture. Exh. H at ¶ 10.

The backbone of the STRmix software system is a computing algorithm called the Markov Chain Monte Carlo (MCMC) method of calculating probable outcomes. *Id*. at 233. The implementation of the MCMC algorithm in STRmix utilizes statistical models to simulate hypothetical true alleles while incorporating stochastic effects. *Id*. It then assesses those simulated alleles and then makes conclusions about what is true DNA as opposed to artifacts in a sample. *Id.* Based on those conclusions, the likelihood ratio is then generated as a further statistical assumption. The reason that MCMC is used is that there are an exponentially enormous number of combinations of assumptions and outcomes that arise from any mixed sample. It would be practically impossible to do such a calculation without a computer running sophisticated software.

ii. How Does STRmix Compare To Other Probabilistic Software Programs?

There is no agreement within the forensic community about which probabilistic software programs or methods to employ when analyzing low template DNA or complex mixture samples. There are at least eight different probabilistic genotyping software programs in the country. Exh. A at 78. Oldman briefly summarizes two others to illustrate their differences with STRmix: The Forensic Statistical Tool (FST), and TrueAllele. These two vary from STRmix and each other in the manner in which they collect data, the necessary assumptions they make to

perform their statistical calculations, and the actual underlying mathematical principles used to make these calculations.

The Forensic Statistical Tool (FST) was developed in-house at the Office of Chief Medical Examiner (OCME) of the City of New York by Dr. Theresa Caragine and Dr. Adele Mitchell and is used in all complex mixture cases in New York City. Similar to STRmix, FST relies on analysts to collect the data used in the calculations and analysis. An analyst reviews the electropherogram and determines whether alleles are present at each locus by utilizing the lab's analytic threshold. The analyst inputs this information, along with a known suspect profile, into the FST software. The analyst then sets the parameters for running the program including, whether the mixture contains two or three contributors. The software then outputs a "Forensic Statistic Comparison Report," summarizing the data that was input and indicating the resultant likelihood ratio. FST differs from STRmix in how it calculates the LR. Unlike STRmix, FST does not use MCMC algorithms in making these calculations, instead relying only on Bayesian statistics. Bayesian statistics describe the probability of an event, based on conditions that might be related to the event. *See* John Butler, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers 459 (Second Ed., Elsevier Academic Press 2005). FST also calculates the "drop-out" rate differently than STRmix. FST calculates the allelic "drop-out" based on the quantitation values of given DNA samples, rather than peak height variation, as STRmix does.

TrueAllele differs significantly from STRmix and FST in the manner in which it collects, interprets, and calculates the data. TrueAllele was developed by Cybergenetics of Pittsburgh, PA under the direction of Dr. Mark Perlin. TrueAllele is a fully continuous probabilistic approach that analyzes the epgs and considers the genotypes at every locus of each contributor, taking into

consideration the mixture weights of the contributors, the DNA template mass, polymerase chain reaction (PCR) stutter, relative amplification, DNA degradation, and the uncertainties of all these variables.

Unlike FST and STRmix, TrueAllele does not rely on an analyst's interpretation of what constitutes a true allele by using analytical thresholds dictated by laboratory protocol in order to collect its data. True Allele instead, considers all the data present in the sample, even those peaks below the lab's analytic threshold. In essence, the calculations made by TrueAllele are based upon more information than used by FST and STRmix. Unlike FST, TrueAllele accounts for "drop-out" rates as a function of peak heights and peak height ratios seen in the sample rather than based on the quantity of DNA in the sample. Like STRmix, but unlike FST, it uses MCMC algorithms to calculate likelihood function that compares genotypes relative to a population and computes a match LR. *See People v. Wakefield*, 47 Misc.3d 850, 859 (Sup. Co. Schenectady Co. Feb. 9, 2015).

iii. The *Hillary* Case: An Example Of How Competing Programs Work With Low Level Mixtures.

People v. Hillary, Ind. No. 15/2015 (Sup. Ct. St. Lawrence County, Aug. 26, 2016) provides an example of how these competing software programs work in practice when analyzing low template mixtures. *See* Jesse McKinley, *Potsdam Boy's Murder Case May Hinge on Minuscule DNA Sample From Fingernail*, The New York Times, Jul. 24, 2016.[30] The *Hillary* case involved the tragic murder of a 12 year old boy in Potsdam, NY in 2011. *Id.* A former local college soccer coach, Oral Nicholas Hillary, was an initial suspect, despite a paucity of

[30] Available at https://www.nytimes.com/2016/07/25/nyregion/potsdam-boys-murder-case-may-hinge-onstatistical-analysis.html.

physical evidence.[31] *Id.* Fingernail scrapings taken from the victim were analyzed first by the New York State police lab. Exh. C (Decision and Order, DNA Analysis Admissibility, *People v. Hillary,* Ind. No. 15/2015 (Sup. Ct. St. Lawrence County, Aug. 26, 2016) at 3. Human analysts there, looking at the electropherograms, determined that almost all of the DNA in the sample was from the decedent, but that there was a trace amount of a second, "minor" contributor present also, possibly from a struggle with the killer. The lab determined that due to insufficient genetic material, Mr. Hillary could be neither exclude nor included as a contributor. Exh. C at 3. NYS police lab then used TrueAllele, which that lab had validated, to see if it could do a better job. TrueAllele produced the same result as the lab's analysts – inconclusive results. *Id;* Exh. D (Notice of Motion To Preclude, *People v. Hillary,* Ind. No. 15/2015 (Sup. Ct. St. Lawrence County, May 31, 2016) at 9. As a result, the prosecution declined to charge Mr. Hillary. *Id.*

In 2014, after running on a campaign to find the killer of Garrett Phillips, the District Attorney indicted Mr. Hillary. Exh. D at 9. The indictment was later dismissed for prosecutorial misconduct in the grand jury. *Id.* Mr. Hillary was reindicted in 2015 and the District Attorney's office asked ESR to test the fingernail scrapings using STRmix. *Id.* at 10. STRmix tested the same sample as the NY state lab and TrueAllele and, contrary to both prior findings, produced an inculpatory likelihood ratio of 10 million. *Id.* Subsequent revisions by ESR resulted in a likelihood ratio of 10,000 and finally, 330,000. *Id.* Subsequent litigation, which ultimately resulted in STRmix being precluded from evidence at the trial, revealed that STRMix, despite coming up with its initial astronomical inculpatory result, had never been validated for this type

[31] Dozens of DNA samples were collected from the crime scene, the body and clothing of the victim, the interior of Hillary's car, Hillary's clothing, and items seized from Hillary's home. Exh. D at 5. Hillary was excluded as a contributor to all samples from the crime scene where comparisons could be made. The victim was excluded from the samples from Hillary's home and vehicle. *Id.* at 8. *See also* Exh. C at 2.-3. The one exception was the fingernail scraping, from which the NY State lab said Mr. Hillary could be neither excluded nor included due to insufficient genetic information. *Id.*

of "extreme" mixture found in the fingernail sample. Exh. C at 10. Mr. Hillary was acquitted at trial.

e. What The PCAST Report Tells Us About The Reliability Of Interpreting Complex Mixtures Using Probabilitic Genotyping Software

As stated above, in 2016, the President's Council of Advisors on Science and Technology (PCAST) issued a Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (Sept. 2016) ("PCAST Report"). The authors were "an advisory group of the Nation's leading scientists and engineers, appointed by the President to augment the science and technology advice available to him from inside the White House and from cabinet departments and other Federal agencies." Exh. A at 3. "PCAST is consulted about, and often makes policy recommendations concerning, the full range of issues where understandings from the domains of science, technology, and innovation bear potentially on the policy choices before the President." The PCAST group includes the President of the Broad Institute of Harvard and MIT, experts in biology, aerospace, astrophysical sciences, natural resources and environment, string and particle theory, electrical engineering and computer science, nanotechnology, and a Medical Doctor. Exh. A at vi.

The PCAST Report was created in response to a 2009 report by the National Research Council that was highly critical of the use, and misuse, of forensics in criminal cases—Strengthening Forensic Science in the United States: A Path Forward ("NRC Forensics Report"). In 2015, President Barack Obama convened some of the country's leading scientists to evaluate whether there were "additional steps on the scientific side," in addition to those already taken in response to the NRC Forensics Report, "to help ensure the validity of forensic evidence used in the Nation's legal system." Exh. A at X. PCAST formed a working group that included several members of the PCAST permanent advisors. In contrast to the 2009 NRC Forensics Report,

which touched on twelve separate disciplines, PCAST examined just six "forensic feature comparison" disciplines: firearms analysis; DNA analysis of single source samples, simple mixture samples, and complex-mixture samples; bite mark analysis; latent fingerprint analysis; footwear analysis; and hair analysis. The group's goal was to determine whether those disciplines were scientifically valid and whether they had a methodology that could be reliably applied—the foundational requirements for admissibility. Exh. A at 1-2.

The group evaluated over 2,000 papers and studies from various sources, including papers submitted in response to PCAST's request for information from the forensic-science stakeholder community. Much like in 2009 NRC Forensics Report, PCAST asked whether each forensic discipline met two key requirements for scientific validity: (1) "foundational validity" – that the method can, in principle, be validly applied; and (2) "validity as applied" – that the method has been reliably applied in practice. Exh. A at 56. To be "foundationally valid," a field must utilize a method that has been subject to "empirical testing by multiple groups, under conditions appropriate to its intended use." *Id*. at 5 (emphasis in original). Those studies must show that the method is "repeatable and reproducible." A method is "repeatable" if, with a known probability, an analyst can reach the same result when analyzing samples from the same sources. A method is "reproducible" if, with known probability, different examiners can obtain the same result when evaluating the same samples. Exh. A at 47. Put slightly differently, a feature comparison method is foundationally valid if studies show it has a "reproducible and consistent procedure" for:

(a) identifying features within evidence samples;

(b) comparing the features in two samples; and

(c) determining, based on the similarity between the features in two samples, whether the samples should be declared to be a proposed identification ('matching rule')." Exh. A at 48.

The studies must also provide "valid estimates of the method's accuracy," in order to demonstrate how often an examiner is likely to draw the wrong conclusions. *Id.* "Without appropriate estimates of accuracy, an examiner's statement that two samples are similar, or even indistinguishable, is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact." *Id.* at 6; *see also id*. at 48 ("Without an appropriate estimate of its accuracy, a metrological method is useless, because one has no idea how to interpret its results."). Simply put, in order to be foundationally valid, the feature comparison method has to "show its work" through studies that document that examiners are able to do what they say they can do, and how often they get the right answer.

Once a method has been established as foundationally valid, to meet the criteria for scientific acceptance, it must also be valid "as applied." A DNA analyst in a given case must be capable of reliably applying the method, and he or she must have actually reliably applied the method in the case at hand. To ensure that the examiner is capable of applying the technique, the field must conduct rigorous proficiency tests evaluating how often an expert reaches the correct answer in circumstances modeling the procedures actually used in case work. *Id*. at 56. To show that the examiner has applied the method reliably in each case, the examiner must make available all procedures used, the results obtained, and any laboratory notes taken. *Id*. Finally, the examiner must make only scientifically valid assertions about the probative value of the identification. The analyst must accurately report the false positive rate for the method, and cannot overstate the significance of his conclusion by making claims that exceed the empirical evidence and the "applications of valid statistical principles to that evidence." *Id*. at 6.

One of the "feature comparison" methods that PCAST evaluated was DNA analysis. PCAST evaluated three separate categories of DNA analysis: (1) single-source samples, (2) simple mixture samples, and (3) complex mixture samples. *Id.* at 69-83. PCAST found that methods for evaluating single source DNA samples and simple, distinguishable mixtures were foundationally valid. *Id*. at 75. PCAST, however, noted a number of problems with complex mixtures, and highlighted real life scenarios in which analysts interpreting complex mixture had come up with wildly different results. One of the scenarios was a 2003 double homicide in which the defendant received the death penalty after the prosecution expert testified the defendant's DNA was a match to a glove and the odds a random person would have matched were 1.1 billion to 1. In fact, further analysis showed that the likelihood was actually 1 in 2 – that is 50% of the relevant population could not be excluded. Exh. A at 77. PCAST further noted that events as recent as 2015 showed that the problems with interpreting complex mixtures were not limited to a few individual cases, but were systemic. *Id.* at 77.

As to probabilistic genotyping software programs, PCAST noted that while they "represent a major improvement over purely subjective interpretation … they still require careful scrutiny to determine (1) whether the methods are scientifically valid, including defining the limitations on their reliability (that is, the circumstances in which they may yield unreliable results) and (2) whether the software correctly implements the methods. This is particularly important because the programs employ different mathematical algorithms and can yield different results for the same mixture profile." *Id*. at 79. Overall, PCAST concluded that the evaluation of complex or indistinguishable DNA mixtures "using probabilistic genotyping software is relatively new and promising approach. Empirical evidence is required to establish the foundational validity of each such method within specified ranges. At present, published

evidence supports the foundational validity of analysis, with some programs, of DNA mixtures of 3 individuals in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum required level for the method." *Id.* at 82.

V. ARGUMENT

a. Legal Standards

Federal Rule of Evidence 702 sets forth the following conditions under which a properly qualified expert can give opinion testimony at trial:

> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based upon sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.

Fed. R. Evid. 702 (emphasis added).

In *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579 (1993), the Supreme Court held that Rule 702 assigns the court a "gatekeeper" role and charges it with the task of ensuring that expert testimony "'rests on a reliable foundation and is relevant to the task at hand.'" *United States v. Hermanek*, 289 F.3d 1076, 1093 (9th Cir. 2002) (quoting *Daubert*, 509 U.S. 579 (1993)). This "gatekeeper" role requires the court to assess "whether the reasoning or methodology underlying the testimony is valid and [] whether that reasoning or methodology properly can be applied to the facts in issue." *Id.* (quotation omitted). The Supreme Court clarified in *Kuhmo Tire, Co. v. Carmichel*, 526 U.S. 137, 147 (1999), "that the district court's duty to act as gatekeeper and to assure the reliability of proffered expert testimony before admitting it applies to all (not just scientific) expert testimony." *Hermanek*, 289 F.3d at 1093.

That is, Rule 702 "establishes a standard of evidentiary reliability" for *all* such matters. *Kuhmo Tire*, 526 U.S. at 149. *See also* Exh. A at 4 (method must have "foundational validity" to be admissible under 702).

Under *Daubert* and *Kumho Tire*, the court must "make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field." *Kuhmo Tire*, 526 U.S. at 152. To that end, *Daubert* and *Kumho Tire* set forth a non-exhaustive list of factors bearing on the reliability and validity of expert testimony: (1) whether the theory or technique can be and has been tested, (2) whether the theory or technique has been subjected to peer review and publication, (3) the known or potential rate of error, (4) whether there are standards controlling the technique's operation, and, (5) the degree of acceptance within the relevant scientific community. *Daubert*, 509 U.S. at 593-94. Moreover, PCAST cautioned that a method is not foundationally valid unless it has "be[en] shown, based on empirical studies, to be repeatable, reproducible, and accurate, at levels that have been measured and are appropriate to the intended application." Exh. A at 4. Judged by these criteria for scientific validity, STRMix's use on Items 12 and 53 is inadmissible under *Daubert*.

b. DNA Results From Items 12 And 53 Should Be Excluded Because Use Of STRmix For Items 12 And 53 Is Not The Product Of Reliable Principles And Methods.

DNA results of Items 12 and 53 do not meet *Daubert's* standards for two reasons: 1) the FBI lab's determination that Items 12 and 53 are four-person mixtures is not reliable and 2) the FBI has not validated testing of mixtures with ratios as extreme as those in Item Nos. 12 and 53. Dr. Dan Krane, a biochemistry Ph.D. and an expert on DNA, wrote a Declaration describing

these concepts in greater detail and explaining his opinion that the FBI's determination that Items 12 and 53 contain no more than four contributors is incorrect. It is attached as Exhibit H.

i. The FBI Lab's Determination That This Is A Four Person Mixture Is Not Reliable.

Determining the number of contributors to a complex mixture appears to be, at best, an art, not a science. While the individual analyst reviews the number of alleles and the peak height ratios to determine the number of contributors, there are no objective standards for doing so, Exh. H at ¶ 11, and the determination of the number of contributors is based on the individual analyst's intuition and experience. The FBI has not provided any proficiency testing or error rates for correctly analyzing the number of contributors for the lab as a whole or any of its DNA analysts including Jaclyn Garfinkle, who did the testing at issue in this case. We know that accurately determining whether a DNA sample contains four, five or six contributors has been tested in the Coble/Bright study, and the results can be fairly described as abysmal. As a result, this technique is not foundationally valid. *See supra,* at 29 (describing "foundational validity as defined in the PCAST report). Even if it was, because we do not have any testing of the analyst at issue in this case, it is also not valid as applied. *Id*. *See also Daubert*, 509 U.S. at 593-94 (the first factor bearing on the reliability and validity of expert testimony is whether the theory or technique can be and has been tested).

In addition, the technique, which is subjective, lacks "standards controlling the technique's operation." *See Daubert*, 509 U.S. at 593-94 (noting that the fourth factor bearing on reliability is whether there are standards controlling the technique's operation). And, as shown by the PCAST report, the technique (to the extent there is one) for accurately determining the number of contributors to four or more person mixtures, is not accepted within the relevant

scientific community. *Id.* (fifth *Daubert* factor). As to the third *Daubert* factor – the potential rate of error – the rate is shockingly high and that, on its own, should preclude use at trial. It bears repeating that the 2018 Coble/Bright study, in which the FBI participated, shows that labs misidentify five-person mixtures two-thirds of the time, and six-person mixtures 100% of the time. Even more worrisome, 58% of five-person mixtures and 69% of six-person mixtures are wrongly interpreted as four-person mixtures. Exh. H at ¶ 14. Considering these error rates, it is simply not possible for the Court to find that the FBI's determination that Items 12 and 53 are four-person mixtures (and thus included within the FBI lab's validation studies) is reliable.

The government may respond that since the PCAST report came out two years ago cautioning that reliable interpretation of complex mixtures is limited to three-person mixtures, the FBI has become competent at accurately interpreting known four-person mixtures. That is beside the point. The FBI's reliability at interpreting known four-person mixtures is only relevant if the FBI can also accurately determine what is a four-person mixture. Considering that a recent study in which the FBI participated showed astronomical error rates at accurately identifying four-person mixtures, that is not a safe bet. And since the FBI has not performed validation studies of five or six-person mixtures using STRmix, the FBI's analysis of Items #12 and #53 do not meet *Daubert's* standards of reliability.

ii. The FBI Has Not Validated Testing Of Mixtures With Ratios As Extreme As Those In Item Nos. 12 and 53.

The PCAST report acknowledged that the "range in which foundational validity has been established is likely to grow as adequate evidence for more complex mixtures is obtained and published." *Id.* However, Items 12 and 53 are so complex that even the validation studies done by the FBI since the PCAST report are not sufficient. As Dr. Krane explained in his Declaration,

mixture ratios investigated in the FBI's internal validation study of STRmix™ v2.4 are less extreme than the mixture ratios inferred by STRmix™ for genotyping results obtained from Items 12 and 53. The FBI validated STRmix up to a 10:1 ratio of the largest to smallest contributor amounts. Item #12 had an 18:1 ratio and Item #53 had a 15:1 ratio. Moreover, the smallest contributor validated by the FBI was 4.8 % while the smallest contributor in Item #12 was 3%. Similarly, the largest contributor validated by the FBI was 58.8% while the largest contributor in Item #53 was 76%. In other words, both Items 12 and 53 were more extreme mixtures than any validated by the FBI.

As a result, the FBI's validation is insufficient because "[t]est samples chosen should represent the spectrum of situations encountered in real-world casework. Profiles representing extreme situations should be included, even if these profiles ultimately might not be interpreted in casework." Hinda Haned, Peter Gill, Kirk Lohmueller et al., Validation of probabilistic genotyping software for use in forensic DNA casework: definitions and illustrations, Science and Justice, 56 (2016) 104-108 at 106. This is because testing must "determine not only when the system works as expected, but also when it may fail. Specifically, it is important to investigate the boundaries of the model within its domain of application. Common characteristics of forensic casework samples that can increase their complexity include multiple contributors, low quantity (provoking possible drop-out) and low quality (e.g., degradation, inhibition, contamination). All of these factors increase ambiguity and reduce information content. Both the limitation of the model and the limitations of the evidence must be tested. *Id.*[32]

[32] Indeed John Butler echoed this criticism in a recent interview. "And here's the challenge: Labs are not prepared to do the complex mixtures. The reality is, all the labs' proficiency tests, as I'm looking at them, are like basic math or algebra. So you're going into a final exam on calculus, but you've only done homework on algebra and basic arithmetic. Are you going to pass that exam? That's the reality of what we're facing." *Putting Crime Scene DNA Analysis on Trial*, (October 11, 2018). https://www.propublica.org/article/putting-crime-scene-dna-analysis-on-trial (accessed November 28, 2018.)

VI. CONCLUSION

Throughout this motion are examples of forensic science gone array. The introduction highlighted the story of William Barnhouse, who was sentenced to 80 years in 1992 based on unreliable hair analysis. But as the rest of this motion showed, unreliable forensic science has found its way to DNA testing too. Oral Hillary was acquitted at trial in 2016 after the judge threw out DNA results that STRmix produced that varied from an inculpatory likelihood ratio of 10 million to 10,000 – and after lab analysts and a separate probabilistic genotyping software refused to make a match. Mr. Winston was convicted and sentenced to death in a 2003 double homicide after the prosecution expert testified the defendant's DNA was a match to a glove and the odds a random person would have matched were 1.1 billion to 1. In fact, further analysis showed that the likelihood was actually 1 in 2 – that is 50% of the relevant population could not be excluded.

Here, almost all the forensic evidence (DNA, trace evidence, fingerprint, shoeprint, chemical) does not inculpate Arapaho Oldman except for possibly two items – Items 12 and 53. These items were analyzed using a process that depends on a fundamental assumption – the number of contributors to a DNA mixture of at least four people – that a 2018 study proves lab analysts get wrong much of the time. Considering this fact, as well as all the other information contained in this motion and the attached exhibits, Arapaho Oldman respectfully requests the Court grant his *Daubert* motion.

DATED this 3rd day of December 2018.

Respectfully submitted,

VIRGINIA L. GRADY
Federal Public Defender

/s/ Galia Amram
Galia Amram
Asst. Federal Public Defender
214 W. Lincolnway Ste. 31A
Cheyenne WY 82001
Telephone (307) 772-2781
FAX (307) 772-2788